THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 6, July–August 1983
Printed in U.S.A.

Human Factors and Behavioral Science:

Retrospective Reports Reveal Differences in People's Reasoning

By D. E. EGAN*

(Manuscript received January 14, 1982)

Reasoning for a class of transitive inference problems was studied and the following questions were experimentally investigated: (1) Can people give reliable retrospective reports about their reasoning processes? (2) Do people who report different reasoning processes actually reason in different ways? (3) Can people be trained to use different reasoning processes? In the situations studied, subjects' retrospective reports about reasoning contained sufficient information to classify the subjects reliably. Subjects classified as using different reasoning strategies made different amounts and different kinds of reasoning errors. As a result of training, subjects could use reasoning processes that they would not have used spontaneously. These results have implications for developing theories of reasoning and for assessing and modifying reasoning-like processes in practical situations.

I. INTRODUCTION

Reasoning, the ability to draw conclusions or inferences from given information, is a prized intellectual talent. Reasoning is known to be associated with successful learning of mathematics.[1] A "reasoning

* Bell Laboratories.

factor" plays a prominent role in reading comprehension.[2] Practical problem-solving tasks such as balancing a checkbook or troubleshooting equipment undoubtedly involve reasoning. Because of its pervasive importance, tests of reasoning have been included in virtually every battery of aptitude tests, and experimental psychologists have studied reasoning extensively in various forms.

Reasoning may play an especially prominent role in the future as more occupations and everyday tasks involve computers. For example, reasoning test scores to some extent predict people's success at computer programming,[3] an occupation that is growing in importance. Reasoning probably is required to use the host of new computer-like devices ranging from automated banking machines and appliance timers to systems for editing and data retrieval. The small bit of "programming" required to activate many new telephone services appears to require mental processes akin to reasoning.

1.1 Three basic questions about reasoning

This paper presents results of basic psychological research concerning three questions about reasoning. The first of these questions is, "Can people give reliable retrospective reports about their reasoning processes?" This question is important, because it asks whether a potentially useful kind of data about the reasoning process meets the scientific standard of repeatability. Using subjective reports to analyze reasoning processes poses much more severe problems of reliability than, say, using meter readings to analyze physical processes.

To be useful data, retrospective reports about reasoning processes must, as a minimum, admit to consistent classification. Reports should contain enough information so that two raters would classify a given report in the same category. Different methods of reporting (e.g., verbal reports and nonverbal reports such as drawings or checklists) should agree with each other so that a reasoner would be classified in the same category no matter which particular method of reporting is used. The classification of reports given by the same person on different occasions should be consistent, as long as other indicators show that the person is using the same reasoning process on those occasions. Retrospective reports must meet each of these requirements to be considered reliable indicators of people's reasoning processes.

A second fundamental question considered here is, "Do people who report different reasoning processes actually reason in different ways?" If different people actually reason in different ways consistent with their reports, two conclusions follow. One conclusion would be that retrospective reports about reasoning are valid evidence about the reasoning process. Reports could be valid simply if people who give a particular kind of report tend to use a particular reasoning process, even if the content of the report does not accurately describe the process. Formally, such reports would have validity because their content would be correlated with performance. A more powerful and interesting kind of validity would require that the content of people's reports accurately reflects how they reason.

Another conclusion would be that data (e.g., reasoning errors) from people giving different kinds of reports do not result from a common underlying reasoning process. To understand reasoning, people using a given reasoning process first may have to be identified and their reasoning data grouped together. Models for the reasoning process appropriate to each group then could be developed.

A very practical aspect of this question is that retrospective reports are often the most easily obtainable data and sometimes they are the only data available in practical situations involving mental processes like reasoning. For example, people's reports about how they do a task or how they use a complex device often are used to evaluate new services or products requiring human interaction. It would be advantageous to have at least one assessment of the validity of these kinds of reports, and to know whether different people are likely to use different reasoning processes spontaneously. The extent to which laboratory studies of reasoning generalize to specific practical situations involving reasoning is not known. However, one benchmark for the validity and variety of retrospective reports about reasoning can be obtained in laboratory studies where it is possible to validate reports.

The third question about reasoning addressed in these studies is, "Can people be trained to use different reasoning processes?" This question introduces a distinction between the reasoning processes people might use spontaneously, and those they could use if trained to do so. In particular, for a specific kind of problem people may be able to follow directions to use a certain reasoning process, even if they would not use that process if left to themselves.

This question has theoretical and practical interest. To the extent that training can induce people to use different reasoning processes, reasoning cannot be conceptualized simply as a stable ability that some people have more of or are better at doing. Instead, a theory of reasoning must explain what factors cause people to discover and use a particular reasoning process when they are competent to use others. Moreover, if people can be trained to reason in different ways, then they probably can be trained to think in various ways about practical tasks involving reasoning. Directions on "how to think about" a task may be an effective aid for learning new procedures (perhaps activating telephone services or interacting with computers) that require reasoning.

1.2 Reasoning problems used in these studies

In all of the studies presented here, people solved reasoning problems involving a simple transitive inference. The problems, known as "three-term series problems," have a common basic form. Examples of the problems are given in Table I.

Three-term series problems were selected for these studies over many other possible kinds of reasoning problems for the following reasons. First, these problems require no specialized content knowledge and therefore seem to be good tools for examining "pure reasoning" uncontaminated by whatever specific information people might know about a problem. Second, the simple verbal structure of problems like those in Table I makes it easy to manipulate certain problem characteristics while controlling others (see below). Finally, three-term series problems have been studied extensively. They have been used on standard tests as markers for deductive reasoning ability[4,5] and much is already known about factors that influence the difficulty of these problems.[6]

Each three-term series problem consists of two premises followed by a question. The premises state a relationship between two of the three terms in the problems. The question concerns the inferred relationship between the remaining pair of terms.

In describing three-term series problems, an important distinction must be made between *relations* and *inverses*. This distinction is based on well-substantiated findings demonstrating that one member of a pair of opposite relational words typically produces better performance than the other member. For example, using the word "above" results in more rapid and accurate performance than the word "below" in a large range of tasks. Similarly, the word "fatter" leads to better performance than "thinner," etc. The psycholinguistic theory of lexical marking[7] may account for some of these differences. For present purposes, an inverse is defined as that member of a pair of opposing relational words that is known to cause the greater difficulty. Conse-

Table I—Sample of the reasoning problems used

Positional Relations	Visual Comparative Relations
Triangle is above circle. Square is below circle. Is triangle above square?	Square is smoother than triangle. Circle is smoother than square. Is triangle smoother than circle?
Circle is left of triangle. Square is left of circle. Is triangle left of square?	Circle is darker than square. Circle is lighter than triangle. Is triangle darker than square?
Square is in back of triangle. Square is in front of circle. Is triangle in front of circle?	Triangle is fatter than circle. Triangle is thinner than square. Is circle thinner than square?

quently, "above" will be referred to as a relation, but "below" will be referred to as an inverse. Similarly, "fatter" is called a relation, but "thinner" is called an inverse.

Having thus defined relations and inverses, the terms of the problems will be distinguished as follows: the "A term" is the initial term in the linear order established by the premises (e.g., when using the relation "rougher" or its inverse "smoother," the A term would be the roughest); the "B term" is the pivot, or middle term; the "C term" is the end term of the linear order (e.g., the smoothest when using rougher/smoother).

Several other details of the problem set are noteworthy. The terms and relationships of each problem were selected to encourage the use of spatial mental representations. The terms were the geometric figures circle, square, and triangle. The relationships involved either positional comparisons (relation/inverse pairs above/below, right of/left of, or in front of/in back of), or nonpositional visual comparisons (rougher/smoother, darker/lighter, or fatter/thinner). For each relation/inverse pair, 16 different problem types were generated. These different problem types resulted from a 2^4 factorial combination of the following factors: (1) the use of a relation or inverse in the premise relating the A term and B term; (2) the use of a relation or inverse in the premise relating the B term and C term; (3) the order of the premises; and (4) the use of a relation or inverse in the problem question. The pattern of reasoning errors made on these different types of problems will be analyzed to test whether various reasoning processes are being used.

II. CAN PEOPLE GIVE RELIABLE RETROSPECTIVE REPORTS ABOUT THEIR REASONING PROCESSES?

2.1 Approach

Regarding the question of the reliability of reports about reasoning, we will consider data from two experiments.[8] In both studies, high school students solved a large number of three-term series problems, and then described their reasoning processes. In the first study, 12 subjects each solved 384 problems. In the second study, 100 subjects each solved 192 problems.

The problems were presented on an audio-tape-playback machine. Each problem began with a speaker saying "Next," then reading the problem, and then pausing five seconds to allow the subject to answer before beginning the next problem. Subjects answered problems by crossing out "Yes" or "No" or "?" next to the number of the problem on an answer sheet. They were not allowed to write anything else. Answers were scored as errors if either the wrong response or a "?" was used.

Several methods of retrospective reporting were used. In the first experiment, subjects were interviewed by the experimenter, and these interviews were tape-recorded and analyzed. In the second experiment, subjects gave written reports—essentially the equivalent of the oral reports in the first study. Following the written reports, subjects then were asked to draw pictures representing how they thought about the problems. As a final method of reporting, subjects were shown two written descriptions of reasoning processes that attempted to capture the two most common kinds of reports found in the first study. Subjects had to choose which of the two descriptions more closely matched their own way of reasoning.

2.2 What people reported

The most striking aspect of subjects' reports was that different subjects claimed to use quite different sets of processes or *strategies* to deal with simple three-term series problems. The difference was most apparent for reports about the visual comparative relations rougher/smoother, darker/lighter, and fatter/thinner. Some subjects claimed to establish an order for the three geometric figures no matter what relation was used. For these subjects, "rougher," for example, would be identified with one end of a vertical or horizontal scale, as would "darker" and "fatter". Subjects described making the transitive inference by mentally arranging the geometric figures roughest to smoothest, darkest to lightest, etc. Other subjects claimed to attribute physical properties to the geometric figures in the case of the visual comparative relations. For example, these subjects described their representation of a rougher triangle as an image of a triangle having a roughly textured surface, or a lighter circle as a picture of a very bright round object. These subjects described making the transitive inference by scanning the images. Distinctions among reports about positional relations were more subtle (see below). Actual examples of written reports are given in Table II.

Reasoning data from subjects who reported similar reasoning strategies were grouped together. The rule for assigning subjects to groups was that if a subject described a representation that clearly involved physical properties for any relation, then the subject was placed in a group labeled "Concrete Properties Thinkers." A subject who claimed to use an ordered mental array for every relation was placed in another group labeled "Abstract Directional Thinkers."

Using this rule, all subjects in the first experiment could be classified into either the Concrete Properties Group ($N = 5$) or the Abstract Directional Group ($N = 7$). In the second experiment, the consensus rating of written reports by two judges identified 18 subjects as

Table II—Examples of written retrospective reports

Abstract Directional Thinkers		
Rougher–Smoother	S#005:	"Rather than imagining a rough/smooth figure, I put the figures in a horizontal line, in my mind, in the order of left/right rather than rough/smooth."
	S#049:	"I pictured the objects in my mind in a line of sequence."
Darker–Lighter	S#003:	"I set up a scale with the lightest on the far right and darkest on the far left and placed the figures on their appropriate spots."
	S#051:	"Placed them in a line up and down, darkest being on top."
Fatter–Thinner	S#086:	"I also used a mental horizontal grid for this relation with the left side of the grid being the 'thin end' and the right side the 'fat end.'"
	S#099:	"Put shapes in order from thinner to fatter."
Concrete Properties Thinkers		
Rougher–Smoother	S#008:	"I also drew a picture, and if something was rough—I would put craters in it in my mind—smooth was just plain white."
	S#098:	"The picture came to mind of corners and smooth edges, then the question was solved."
Darker–Lighter	S#062:	"In my mind, I 'colored in' the object that was darkest."
	S#080:	"I listened to the problem and tried to solve it mentally, at times picturing the objects colored in or not."
Fatter–Thinner	S#022:	"This (fatter/thinner problem) was hard. I had to think of the shapes as squeezed or pulled."
	S#100:	"Made them (the figures) fatter and thinner in my head."

Concrete Properties Thinkers, and 42 as Abstract Directional Thinkers.

The classification rule identifying a subject as a Concrete Properties Thinker on the basis of a single concrete report was motivated by simplicity. Subsequent analyses suggest that the rule, while admittedly crude, did manage to separate people into two groups that used a particular set of reasoning processes fairly consistently. First, subjects who reported a concrete representation for one relation were very likely to report using such a representation for other relations. For example, 17 of 18 subjects identified as Concrete Properties Thinkers in the second study gave reports having a concrete representation for two or more relations. Second, in the statistical analysis of reasoning errors, reasoning groups did not interact with the different relation/inverse pairs, suggesting that each group consistently used one reasoning process. Third, a key-word analysis of the written retrospective reports suggests that the two groups also handled positional relations differently. People classified as Concrete Properties Thinkers used more words in their reports that suggest the use of a visual image (variants of the words "picture" and "draw") for positional problems (e.g, "I pictured the objects in a row"). People classified as Abstract Directional Thinkers used more words suggesting the use of an order-preserving scale (variants of the words "put", "order", "line," and "horizontal/vertical") for positional problems (e.g., "I put the objects in order on a horizontal line"). This interaction of key-word types and reasoning groups was statistically reliable.

To summarize, the majority of people reported one of two sets of reasoning processes or strategies for the three-term series problems. Some people claimed to preserve the information in the premises of at least some problems by means of an image capturing the visual features or stated position of the geometric terms. Others claimed to preserve the information in all premises by means of an abstract ordering of the geometric terms.

2.3 Reliability of retrospective reports

Several methods were employed to assess various aspects of the reliability of the retrospective reports. The first, alluded to earlier, was to assess the agreement between two different judges who categorized subjects on the basis of their written reports in the second experiment. Each judge classified the 100 subjects as Abstract Directional, Concrete Properties, or Other/Not Clear. Table III shows that the two judges agreed on the classification of 82 percent of the subjects. Almost all cases of disagreement occurred when one judge classified a subject as using one of the two identified strategies, but the other judge classified the subject as using an Other/Not Clear strategy. Compared to several additional studies[9] the consensus shown in Table III is the "worst case." Other estimates of interjudge agreement have ranged up to 95 percent.

A second analysis assessed the agreement among different methods of reporting in the second experiment. The pictures drawn and forced-choice strategy selections made by subjects were compared to the classification of their written reports. Pictures were classified as indicating the Concrete Properties strategy if they depicted geometric objects with altered physical properties (e.g., a pockmarked surface

Table III—Classification of written reports by two judges (Experiment II)

	2nd Judge's Categories		
1st Judge's Categories	Concrete Properties	Abstract Directional	Other/Not Clear
Concrete Properties	18	0	6
Abstract Directional	1	42	5
Other/Not Clear	3	3	22

depicting "rougher," shading depicting "darker," etc.). Pictures showing horizontal or vertical orderings of rather standard geometric figures were classified as indicating the Abstract Directional strategy. The classification of drawings agreed with the classification of written reports for 56 of the 60 subjects (93.3 percent) whose written reports had been classified by consensus as Abstract Directional or Concrete Properties. The analysis of forced-choice strategy selections showed that 51 of the 60 subjects (85 percent) chose the strategy description consistent with the classification of their written report.

To assess the long-term stability of reported reasoning strategies, 38 subjects who participated in the second experiment and who had been classified by consensus as Abstract Directional or Concrete Properties Thinkers were recalled six months later for another study. After solving some warm-up problems, subjects gave written reports describing their reasoning strategies. These reports were classified in the previously described manner and this classification was compared to the classification of the subjects performed six months earlier. The results in Table IV indicate that subjects' reports about reasoning have some, but not perfect, stability over time and across different presentation conditions (the former reports were given after listening to problems, the latter after reading problems). The stability of verbal reports estimated by the four-fold point correlation based on Table IV is $r = 0.59$ ($p < 0.01$). Specifically, 95-percent of the subjects earlier classified as Abstract Directional again reported that strategy, but only 59 percent of the original Concrete Properties Thinkers reported that strategy again six months later. The instability of the latter group may have been due to unreliable reports or classification procedures on one hand, or actual changes in reasoning strategies[9] on the other.

2.4 Summary

The reliability of retrospective reports about reasoning has been established in that: (1) different judges show considerable agreement on how to classify reports, (2) the classification of written reports agrees to a large extent with classifications based on other nonverbal methods of reporting, and (3) the classification of reports given by

Table IV—Number of subjects using strategies initially and six months later

Strategy Used Initially (Experiment II)	Strategy Used Six Months Later	
	Abstract Directional	Concrete Properties
Abstract Directional	20	1
Concrete Properties	7	10

people at different times has some stability. On the other hand, retrospective reports about reasoning or perhaps the present procedures for classifying reports are not perfectly reliable. Compared to paper-and-pencil tests having finely graded scores, the reliability of retrospective reports is somewhat low. In particular, the classification of reports given at different times and under different conditions of presenting problems is not always the same. Despite these difficulties, the great majority of retrospective reports about reasoning in these studies contain sufficient information to be classified consistently. This is a necessary condition for reports to be useful in exploring the process of reasoning.

III. DO PEOPLE WHO REPORT DIFFERENT REASONING PROCESSES ACTUALLY REASON IN DIFFERENT WAYS?

3.1 Approach

Reasoning errors from the two studies previously described will be used to analyze the validity of retrospective reports about reasoning and to gain further understanding of reasoning processes. Error data from the two groups of subjects reporting different reasoning strategies will be compared at successively finer levels of detail. The overall error rates from the two groups will be analyzed first. Then, general patterns of interaction in the error data will be discussed. Next, the effects of specific problem factors hypothesized to affect a particular reasoning process will be tested. Finally, two models of the process of making a transitive inference will be described and tested. The goal of this section is to demonstrate that different models of the reasoning process are required to account for reasoning errors made by the two groups of subjects who reported different reasoning strategies.

3.2 Differences in reasoning errors between groups reporting different strategies

The first attempt at assessing the validity of retrospective reports asked whether the overall reasoning error rate was different for people giving different reports. In the two studies described above, subjects giving Abstract Directional reports made significantly fewer errors than those giving Concrete Properties reports. In the first study, Abstract Directional Thinkers had an error rate of 10.3 percent compared to the Concrete Properties Thinkers' error rate of 38.0 percent. In the second study, the corresponding error rates were 21.0 percent and 27.9 percent. This difference in error rate favoring the Abstract Directional subject now has been found repeatedly.[10] One interpretation of this result is that the Abstract Directional strategy is more efficient for the transitive inference problems used in these studies.

A further study[9] required subjects to describe their reasoning processes at a number of points in a lengthy sequence of three-term series problems. In that study, a change in the strategy reported by a subject was accompanied by a corresponding change in the reasoning error rate. Thus, if a subject reported shifting from the Concrete Properties strategy to the Abstract Directional strategy, the subject's performance improved. For subjects reporting no shift in reasoning strategy, reasoning performance was fairly stable at a low or high level, depending on the strategy reported. This study provides evidence of the validity of different reports given by the same subject. It also suggests why reports by some subjects changed after six months in the previous study: the subjects' reasoning processes may have changed over time.

People reporting different reasoning strategies not only exhibited different overall levels of reasoning errors, but they also exhibited different patterns of reasoning errors. This fact is demonstrated in a general way by a statistically reliable interaction between Report Groups and Problem Types found in the two original experiments.[8] This interaction means that factors causing reasoning problems to be more or less difficult in one group of reasoners were not the same as the factors causing difficulty in the other group. People who gave different reports made different amounts and different kinds of reasoning errors.

Thus far, the validity of reports has been assessed in a formal but rather indirect way. At the next level of detail, we might ask whether the patterns of reasoning errors made by subjects are consistent with the reasoning process subjects claim to be using. Consider reports of Abstract Directional Thinkers who claim to construct an ordered mental array of the geometric terms used in these problems. Previous theories[11,12] have asserted that it should be easier to construct a direct spatial array from the ends toward the middle, rather than from the middle outward. This so-called "end-anchoring principle" leads to a prediction regarding the difficulty of solving various types of three-term series problems. For Abstract Directional Thinkers, reasoning errors in a problem should be directly related to the number of premises that have the middle or pivot term stated first. For people using the Concrete Properties strategy, the end-anchoring principle should be irrelevant.

This prediction was confirmed by patterns of reasoning errors. In both studies, reasoning error rate for Abstract Directional Thinkers was a monotonic function of the number of premises in a problem that began with the middle or pivot term. This factor accounted for highly significant amounts of the variance in the difficulty of different types of problems for Abstract Directional Thinkers (82.2 percent of the variance in Experiment I, 71.8 percent in Experiment II). Data for

Concrete Properties Thinkers were quite different. Reasoning error rate was not a monotonic function of the number of pivot-first premises in either experiment, and this factor accounted for much smaller amounts of the variance in problem difficulty for this group of subjects (17.6 percent in Experiment I, and 20.3 percent in Experiment II). This analysis suggests that Abstract Directional Thinkers are constructing a mental array as they report, and that Concrete Properties Thinkers are reasoning in a different way.

Other analyses of specific problem factors[8] have found that reasoning errors by Concrete Properties Thinkers depend on the number of inverses used in a problem (e.g., using words like "smoother" and "thinner"), as well as the number of times a relation and inverse are alternated in the statement of a problem. Inverses cause extra difficulty for Concrete Properties Thinkers because concrete representations of a property like smoothness may not be as easy to generate as concrete representations of a property like roughness. Abstract Directional Thinkers were less sensitive to such factors. The most difficult kind of problem for Concrete Properties Thinkers was one with premises like, "Circle is rougher than square. Circle is smoother than triangle." In such problems, the Concrete Properties Thinker presumably imagines first a rough circle next to a square, and then imagines a smooth circle next to a triangle. Such problems are difficult to answer when this inconclusive image is scanned.

3.3 Two models of reasoning

Models attempting to capture the reasoning process used by Abstract Directional and Concrete Properties Thinkers are presented in Tables V and VI, respectively. These models try to give a coherent account of the patterns of reasoning errors and the kinds of reports given by subjects. Each model contains parameters representing hy-

Table V—Model for abstract directional thinkers

Process	Problem Factors	Model Parameters
1. Encode Premise 1		
2. Establish abstract scale		
3. Arrange first two terms placing grammatical subject first on scale		
4. Encode Premise 2		
5. Find third term	Is third term grammatical subject or object?	SEARCH
6. Position third term	Does third term fall in "Natural" next position?	POSITION
7. Encode question		
8. Scan the scale		
9. Respond		

Table VI—Model for concrete properties thinkers

Process	Problem Factors	Model Parameters
1. Encode Premise 1		
2. Generate Image Pair 1 by assigning property to grammatical subject	Is difficult (inverse) relation used?	GENERATE
3. Encode Premise 2	Is relation the same as that in #1?	ENCODE
4. Generate Image Pair 2 by assigning property to grammatical subject	Is difficult (inverse) relation used?	GENERATE
5. Encode question	Is relation the same as that in #3?	ENCODE
6. Scan images	Are the images conclusive?	SCAN
7. Respond		

pothetical mental processes that are executed various numbers of times depending on the structure of a specific type of problem.

3.3.1 The abstract directional model

Abstract Directional Thinkers (see Table V) are assumed to encode the first premise and establish a mental scale for a problem. Then, the two terms stated in the first premise are arranged on the scale, the grammatical subject being placed first. The second premise is then encoded, and the subject searches for the third, or missing, term. This search is easier if the third term is the grammatical subject rather than the object of the second premise. The value of the SEARCH parameter (0 or 1, respectively) reflects this difficulty, and accounts for the effect of starting the second premise with the pivot term. Next, the third term is positioned on the mental scale, and it is assumed that there are three distinct cases for this operation. The easiest case (POSITION = 0) occurs when the third term is placed next in the sequence established by the first two terms. For example, if the first two terms are arranged smooth → rough, positioning the third term is easiest if it is roughest. If the first two terms are placed rough → smooth, then positioning the third term is easiest when it is the smoothest. Two more difficult cases exist and correspond to problems beginning with a pivot-first premise. In the easier of these cases (POSITION = 1) the third term does not fall next in sequence, but instead it must be placed at the end of the scale associated with the relation. In the remaining, most difficult case (POSITION = 2), the third term again does not fall in sequence, but must be positioned at the end of the scale associated with the inverse.

3.3.2 The concrete properties model

The model for Concrete Properties Thinkers suggests that these subjects generate and compare images of objects having the stated properties. For each premise, Concrete Properties Thinkers (Table

VI) are assumed to encode the premise and then generate an image pair in which the grammatical subject takes on the property stated in the premise, while the grammatical object remains neutral. After two such pairs have been generated, the question is encoded, and then the two image pairs are scanned for the answer. Differences in difficulty among problems are assumed to arise from three sources, each corresponding to a parameter in the processing model for Concrete Properties subjects. One kind of difficulty has to do with whether the relation or inverse is used in each premise. Using an inverse presumably makes the appropriate image pair more difficult to generate. For a given problem, the parameter GENERATE takes on a value equal to the number of difficult images required (0, 1, or 2). The parameter ENCODE reflects the difficulty of alternately accessing a relation and its inverse. This parameter equals the number of alternations between a relation and inverse as a problem is read (0, 1, or 2). Finally, the parameter SCAN reflects the difficulty of dealing with images that are inconclusive. As noted previously, problems in which the B term takes on a property in one image pair and then takes on the inverse property in the other pair are especially difficult for Concrete Properties Thinkers. Such problems produce inconclusive image pairs in which the A and C terms are both neutral. Confronted with this type of problem, Concrete Properties Thinkers may guess or reformulate one of the premises to arrive at an answer. The SCAN parameter has the value 1 for such problems and 0 otherwise.

3.4 Comparing models to data

The two models were compared to the data of Abstract Directional and Concrete Properties Thinkers from each experiment. The proportion of variance in problem difficulty uniquely associated with each parameter in each model was determined by stepwise multiple regression. The data are shown in Table VII. For both experiments, the Abstract Directional model was the better predictor of performance for Abstract Directional Thinkers, while the Concrete Properties model was the better predictor for Concrete Properties Thinkers. If errors on the various problem types are combined across experiments, the Abstract Directional model accounts for 90.3 percent of the variance in problem difficulty for Abstract Directional Thinkers (the Concrete Properties model accounts for 80.4 percent of the variance for this group). Both the SEARCH and POSITION parameters account for significant and unique portions of variance in problem difficulty for Abstract Directional Thinkers. The Concrete Properties model accounts for 80.1 percent of the variance in problem difficulty for the combined Concrete Properties Thinkers (the Abstract Direc-

Table VII—Proportion of variance* in problem difficulty attributable to parameters of two models

	Experiment					
	I		II		I + II	
Parameters	Abstract Group	Concrete Group	Abstract Group	Concrete Group	Abstract Group	Concrete Group
Abstract Directional Model						
1. SEARCH	0.576†	0.021	0.668†	0.105	0.668†	0.076
2. POSITION	0.288†	0.305‡	0.208†	0.201	0.235†	0.311‡
$\sum R^2$	0.864†	0.326	0.876†	0.306	0.903†	0.387‡
Concrete Properties Model						
1. SCAN	0.794†	0.372‡	0.724†	0.374‡	0.768†	0.474†
2. GENERATE	0.001	0.185‡	0.016	0.166‡	0.011	0.221†
3. ENCODE	0.024	0.079	0.024	0.087	0.025	0.106‡
$\sum R^2$	0.819†	0.636†	0.764†	0.627†	0.804†	0.801†

* These proportions are increments in R^2 values due to each parameter. The order in which the parameters are given corresponds to the step at which they entered the regression equation for the group of subjects appropriate to a particular model.

† $p < 0.01$

‡ $p < 0.05$

tional model accounts for 38.7 percent of the variance for this group), and each of the parameters SCAN, GENERATE, and ENCODE accounts for significant and unique variance.

Two aspects of the results of the model-fitting procedure should be clarified. First, the reliability of the error rates on the 16 problem types imposes a theoretical upper limit on the amount of variance for which any model can account. Therefore, it is important to estimate the data's reliability and compare that estimate to the R^2 of the best-fitting model.

Reliability estimates suggest that it would be difficult to improve the fits of the models appropriate to each group of subjects in Table VII. The estimated reliability of the Abstract Directional data combined across experiments was 0.951, so the Abstract Directional model ($R^2 = 0.903$) accounted for 0.903/0.951 or 95.0 percent of the reliable variation in the Problem Type data for those subjects. The fit of the Concrete Properties model ($R^2 = 0.801$) actually slightly exceeded the theoretical upper limit of the reliability of the combined Concrete Properties data (estimated reliability was 0.727).

Second, it is important to note that certain parameters of the two models are correlated in the 16 problem types used. The most important example of this confounding occurs for the parameter SCAN in the Concrete Properties model, which is correlated with both the SEARCH and POSITION parameters in the Abstract Directional model. These correlations account for the contribution of SCAN to variance in problem difficulty for Abstract Directional subjects. This interpretation of the contribution of SCAN is consistent with the fact that it is the only parameter in the Concrete Properties model that correlates with performance for Abstract Directional subjects, and that the two-parameter Abstract Directional model accounts for more variance in that group than the three-parameter Concrete Properties model.

3.5 Summary

Retrospective reporting under the conditions studied here apparently is one case in which reports about reasoning processes contain valid information. The fact that the two groups of people who reported different reasoning strategies actually reasoned in different ways is supported by (1) the different overall levels of reasoning errors made by the two groups, (2) the different general patterns of reasoning errors made by the groups, (3) the differential effects of specific problem factors hypothesized to influence difficulty in one group but not the other, and (4) the fits of different process models of reasoning to the reasoning error data of the two groups.

IV. CAN PEOPLE BE TRAINED TO USE DIFFERENT REASONING PROCESSES?

4.1 Approach

A further study[9] dealt with the question of whether people can be trained to use different reasoning processes. In that study, 65 adult women solved a small number of three-term series problems and reported their reasoning processes. As in the experiments described previously, these reports identified the reasoning strategies that subjects spontaneously used. The subjects were then randomly assigned to two groups. One group received training in applying the Abstract Directional strategy to a new set of three-term series problems. The other group was trained to apply the Concrete Properties strategy to the same problems. Reasoning errors made by the two groups of subjects after receiving training were compared.

4.2 Training in reasoning strategies

The training consisted of short descriptions of the models in Tables V and VI and examples showing how to apply them. The training was tailored specifically to a new set of problems involving the relation/inverse pair happy/sad. The terms for these problems were the names of three imaginary people, "Rich", "Dot", and "Harry". A typical problem was therefore, "Rich is happier than Dot. Harry is sadder than Dot. Is Harry happier than Rich?"

Subjects in the Concrete Properties training group were told to represent premises by vividly imagining faces having different features. Illustrations of the faces were drawn such that the people's names suggested the image of the correct face. Thus, "Rich" was depicted as a man wearing an expensive top hat, "Dot" was drawn with freckles, and "Harry" was pictured with a beard and mustache. Subjects were told to represent each premise by visualizing a pair of faces in which the face of the grammatical subject was smiling or frowning, depending on the wording of the problem. The two pairs of images then were to be scanned to answer the question for each problem.

Subjects given the Abstract Directional training were told to imagine a scale with "Sad" on the left and "Happy" on the right, and to place the names of the people on the scale appropriately as a problem was read. The order of the names on the scale then was to be used to make the transitive inference required to answer the question.

Following training, subjects solved 32 happy/sad problems. Subjects next rated the difficulty of applying the strategy they were trained to use. The rating scale ranged from 1 (extremely easy to use the strategy) to 6 (extremely difficult to use the strategy). After giving these ratings, subjects described the strategy they would have used if they had not received training.

4.3 Results of strategy training

Reasoning errors made by the two training groups were compared at successively finer levels of detail. The analyses parallel those applied previously to errors from subjects giving retrospective reports of spontaneously adopted strategies.

First, the group trained to use the Concrete Properties strategy had a significantly higher overall error rate (34 percent) than the group trained in the Abstract Directional strategy (17 percent). Second, the two training groups exhibited different general patterns of reasoning errors, as indicated by a statistically reliable interaction of training groups and problem types. Third, the specific problem factors found to distinguish the spontaneous report groups had analogous effects in the training groups. Errors made by the Abstract Directional training group were related strongly to the number of premises in a problem beginning with the pivot term (the "end anchoring effect" described previously), but were not strongly related to the number of inverses in a problem (uses of the word "sadder"). Subjects trained in the Concrete Properties strategy tended to show the complementary pattern. Fourth, when the process models in Tables V and VI were fitted to the error data of each group of subjects, the appropriate process model gave the better fit in each case (see Table VIII).

Two further analyses related the strategy reported by subjects prior to receiving training to their performance after training. In the first of these analyses, subjects were grouped by the strategy they reported prior to training, and the reasoning errors of the different groups were

Table VIII—Proportion of variance* in problem difficulty attributable to two strategy models

Parameters	Abstract Directional Group	Concrete Properties Group
Abstract Directional Model		
1. SEARCH	0.691†	0.330‡
2. POSITION	0.003	0.198
$\sum R^2$ (Percent of Reliable Variance)	0.694† (93.9%)	0.528† (57.4%)
Concrete Properties Model		
1. SCAN	0.372‡	0.743†
2. GENERATE	0.067	0.113‡
3. ENCODE	0.014	0.012
$\sum R^2$ (Percent of Reliable Variance)	0.453 (61.3%)	0.868† (94.4%)

* These proportions are increments in R^2 values attributable to each parameter. The order in which parameters are given corresponds to the step at which they entered the regression equation.
† $p < 0.01$
‡ $p < 0.05$

compared (see Table IX). The general result was that subjects trained to use the Abstract Directional strategy made fewer errors than those trained to use the Concrete Properties strategy no matter what strategy was initially reported. This result suggests that people can be trained to use a particular reasoning strategy even if they have not adopted that strategy spontaneously.

Subjects given Concrete Properties training rated their strategy significantly more difficult to use ($\bar{x} = 3.81$) than those given Abstract Directional training ($\bar{x} = 2.36$). These ratings then were related to the strategy reported by subjects prior to training. All subjects given Abstract Directional training found that strategy relatively easy to use no matter what their initial strategy had been. For people trained to use the Concrete Properties strategy, the results were different. The Concrete Properties strategy was rated more difficult to use by Abstract Directional Thinkers ($\bar{x} = 4.41$) than by Concrete Properties Thinkers ($\bar{x} = 3.20$). Virtually all the subjects who initially reported using the Abstract Directional strategy indicated that they would have applied that strategy to the happy/sad problems if there had been no training. This pattern of results suggests that people can appreciate a good reasoning strategy: subjects rated a less efficient reasoning strategy as "difficult to use," especially if they knew a more efficient strategy.

4.4 Summary

People can be trained to use different mental processes for making transitive inferences in three-term series problems. This fact has been demonstrated by analyses of the reasoning errors made by people after receiving training in different strategies. Results closely parallel the results of previous analyses of errors made by people retrospectively reporting the different reasoning strategies. Further results suggest that training in a particular strategy can be effective even for those people who did not report the strategy in spontaneous reasoning. People also appear to be sensitive to the difficulty of using various reasoning strategies, and rate a less efficient strategy "difficult to use," especially if they know a better one.

Table IX—Reasoning error rates made by subjects after strategy training

	Strategy Reported Prior to Training		
	Abstract Directional	Concrete Properties	Other/Not Clear
Strategy trained to use	$\bar{X}$ (N)	$\bar{X}$ (N)	$\bar{X}$ (N)
Abstract Directional	0.10 (12)	0.09 (6)	0.25 (15)
Concrete Properties	0.33 (11)	0.33 (5)	0.36 (16)

V. GENERAL DISCUSSION

5.1 Retrospective reports about reasoning

These studies provide a direct test of the reliability and validity of retrospective reports about reasoning. Generally, reports by different subjects contained sufficient information to classify the subjects consistently. The content of subjects' reports also was related systematically to the patterns of reasoning errors subjects made.

While retrospective reports proved very useful in these studies, they also had definite limitations. The classification of different people on the basis of their reports was not perfectly reliable. Another limitation was that a sizable minority of people gave reports that were either idiosyncratic or incoherent (Other/Not Clear reports). Subjects also reported metaphors or general descriptions of reasoning strategies (see Table II) rather than complete and detailed models like those in Tables V and VI. Using a crude classification rule, the majority of subjects' reports could be grouped reliably into two categories, and people giving different kinds of reports tended to exhibit different amounts and patterns of reasoning errors.

The fact that different people reported and used different reasoning processes has two practical implications. One is that it may be possible to obtain useful reports of reasoning-like processes in practical situations where a procedure or device is to be evaluated. Reports may lead to redesigning tasks to make them more compatible with people's reasoning processes. Another implication is a methodological suggestion. If reports on reasoning-like processes are used, it may be wise to obtain them from a large number of different people to gauge the range of mental processes people are likely to adopt.

The general conditions under which retrospective reports about thought processes are reliable and valid remain to be established. The method used in these studies was to have subjects carefully describe how they thought through specific kinds of problems immediately after attempting to solve a large number of the problems. Subjects were not asked why they chose a particular strategy, a kind of judgment that people are notoriously poor at making.[13] Subjects also were not asked to "think out loud" while solving problems, a technique that might have yielded more precise descriptions of the reasoning process. The cost of that technique in the present experiments on reasoning is that it would have prevented the collection of unbiased reasoning error data, and therefore would have clouded the test of the validity of reports. For other purposes, the technique of "thinking out loud" may be quite acceptable. What kinds of mental processes are reportable and which techniques are best for reporting them are two questions that must be answered before reports about thought processes generally can be used with confidence.[14]

5.2 Differences in people's reasoning

An important result of these studies is that different people spontaneously adopted different reasoning strategies for solving very simple, highly stereotyped, three-term series problems. Different models of reasoning processes accounted for large amounts of the variance in problem difficulty for people classified as using different reasoning strategies. These models also received support from results of the training study in which people were directed to use one reasoning strategy or the other (Table VIII). While the process models provide a good first approximation to the reasoning processes employed by different people, the more basic and important result is that people spontaneously reason in different ways.

Training has been shown to cause people to adopt different reasoning processes in a rather simple way: people can follow different sets of directions on how to reason. Subsequent studies[9] identified two other factors that influence in a subtle way whether people use one reasoning strategy or another. One factor is aptitude for visualizing spatial transformations of figures. People good at spatial visualization are more likely to adopt the Abstract Directional strategy spontaneously than those who have difficulty with spatial visualization. The use of different reasoning strategies for three-term series problems is not, however, influenced by verbal aptitude. A second factor influencing the adoption of a reasoning strategy is the context in which reasoning problems are posed. The strategy used for a particular problem is influenced by surrounding problems. Some context problems apparently suggest or allow the development of good strategies while others inhibit strategy development.

A new theory of reasoning is required to account for differences in people's reasoning. The emerging picture includes a dynamic and modifiable process in which reasoning strategies are developed. This phase of strategy development may be influenced by people's basic capacities, the context in which the reasoning occurs, and training. The fact that people can rate the difficulty of using different reasoning processes suggests that feedback of this kind also may be involved in strategy development. Certainly this picture is a far cry from the idea that reasoning is a fixed ability in which people differ by the amount they have or how well they can perform. The emerging picture holds the hope that reasoning processes used by different people can be understood, and that the understanding might lead to redesigning some tasks that many people find very difficult.

5.3 Training in reasoning

The training study reviewed here demonstrates that people can learn to use different reasoning processes with consequent effects on their reasoning performance. A good characterization of the Abstract Directional training would be that it suggested an efficient general approach that most subjects could use, but that some subjects would not have discovered spontaneously. This result tends to confirm informal observation that a suggested strategy for dealing with a complex problem (e.g., thinking of a queueing problem in terms of a "pushdown stack", adopting spatial metaphors for programming problems, etc.) can be very helpful and will lead to certain patterns of performance. Perhaps there are many situations in which directions on "how to think about" a task (e.g., activating a telephone service or interacting with a computer) might reduce errors and lead to more predictable patterns of performance compared to directions in which learners must develop a strategy on their own.

The training study did not address the very important question of whether people can be trained to reason better in general. It is unlikely that the effects of strategy training for a specific simple transitive inference problem would generalize to cause subjects to reason more efficiently in many other situations. People who in general are good reasoners probably have discovered and used a large set of strategies that they can retrieve in given situations. Because of their basic capabilities (e.g., good spatial visualization) and previous experience, good reasoners probably are also very good at generating new alternative strategies for novel problems. The results reviewed here suggest that people can be trained to deal more effectively with some specific situations involving reasoning. This basic fact certainly does not reduce the likelihood of finding ways to train people to reason better in a large range of situations.

VI. ACKNOWLEDGMENTS

Dorothea Grimes-Farrow collaborated on the first two studies reviewed here. Peter Meany and Anthony Rooney assisted in collecting and analyzing data for several of the studies. E. Z. Rothkopf gave many helpful suggestions for revising previous drafts of this paper.

REFERENCES

1. V. A. Krutetskii, *The Psychology of Mathematical Abilities in School Children*, J. Teller, Trans; J. Kilpatrick and I. Wirszup, Eds., Chicago: The University of Chicago Press, 1976.
2. F. Davis, "Psychometric research on comprehension in reading," Reading Res. Quart., *7* (Summer 1972), pp. 628–78.
3. D. M. Irons, "Cognitive correlates of programming tasks in novice programmers," Proc. of Human Factors in Comp. Syst., Gaithersburg, Md., March 15–17, 1982.
4. C. Burt, "The development of reasoning in school children," J. Exp. Pedagogy, *5* (1919), pp. 68–77.
5. L. L. Thurstone, "Primary mental abilities," Psychometric Monogr., *1*, 1938.
6. R. J. Sternberg, "Representation and process in transitive inference," J. Exp. Pscyh.: Gen. *109* (June 1980), pp. 119–59.

7. H. H. Clark, "Influence of language on solving three-term series problems," J. Exp. Psych., *82* (November 1969), pp. 205–15.
8. D. E. Egan and D. D. Grimes-Farrow, "Different mental representations spontaneously adopted for reasoning," Mem. & Cog., *10* (July 1982), pp. 297–307.
9. D. E. Egan, unpublished work.
10. D. E. Egan, unpublished work.
11. C. B. DeSoto, M. London, and S. Handel, "Social reasoning and spatial paralogic," J. Pers. and Soc. Psych., *2* (1965), pp. 513–21.
12. J. Huttenlocher, "Constructing spatial images: a strategy in reasoning," Psych. Rev., *75* (November 1968), pp. 550–60.
13. R. E. Nisbett and T. D. Wilson, "Telling more than we can know: verbal reports on mental processes," Psych. Rev., *84* (May 1977), pp. 231–59.
14. K. A. Ericsson and H. A. Simon, "Verbal reports as data," Psych. Rev., *87* (May 1980), pp. 215–51.

AUTHOR

Dennis E. Egan, A.B., 1969 (Psychology), College of the Holy Cross; M.A., 1973 (Mathematics), Ph.D., 1973 (Experimental Psychology), University of Michigan; Naval Aerospace Medical Research Laboratory, 1973–1976; Bell Laboratories, 1976—. Mr. Egan has worked on individual differences in cognitive abilities, especially spatial ability and reasoning. At Bell Laboratories, he has explored ways of adapting instruction to individual differences in abilities, experience, and other learner characteristics. Mr. Egan is currently a member of the Learning and Instruction Research Department.